SRAM Memory Components Operating at Sub-threshold Voltages

J. F. Ryan, J. Huang, D. Unluer
ECE 563 – Fall 2006
University of Virginia
< jfr7p,jh3wn,du7x>@virginia.edu

ABSTRACT

This paper compares the performance of the common 6T SRAM design and an improved 8T design in sub-threshold operation. Optimal parameters for the transistors are derived based on experimental data. The impact of process variation on SRAM's performance and SNM are analyzed, together with redundancy as an alternative approach to reduce failure rate. Finally, in addition to bit-cells, various types of sense amplifiers operating in the sub-threshold region are compared using metrics of resolution speed and input-referred offset voltage. Ways to improve these metrics such as body-biasing and increasing output impedance were also examined.

1. INTRODUCTION

In the past a few years, a number of new SRAM bit-cell topologies have been proposed to provide better performance while keeping overhead small. In this paper, we contrast the common 6T model with an improved 8T model in performance and reliability.

As power becomes a huge issue in today's IC industry, SRAM memories operating at sub-threshold voltages are gaining more and more attention, considering memory related circuitry constitutes about 50% of the total power consumption in a microprocessor. While some of the previous issues around memory designs are still present (such as transistor sizing), the new area of sub-threshold operation gives rise to brand new issues. Process variation is particularly troublesome due to deeper scaled transistor geometry and a nonlinear relationship between V_T and performance. SNM (Static Noise Margin) is an important metric of reliability and we will focus on the Read-SNM in particular. Since SNM varies significantly with process variation, failure of certain memory cells may be possible. We observe a basic trend of increased the sizes versus the Read-SNM and failure rate. As an alternative, we can use redundant rows and columns to correct the bad cells. A trade-off between area overhead and failure rate is evaluated as well as the advantage and disadvantage of two solutions (redundancy vs. scaling). Monte Carlo simulation is used to imitate the effect of process variation.

2. BIT-CELL TOPOLOGIES

The area of a SRAM bit-cell is needed to be as small as possible, in order to maximize the amount of memory on a chip. As the area of the bit-cells decreases, process variations' affect on the access speed of the cells becomes a bigger problem. [1] Because of process variations, bit-cells become unstable and unusable (reflecting the wrong data), and so they need to be replaced by redundant rows and columns. [2] [3] These redundant rows and columns also occupy a large area on the chips, and thus the best way to maximize the memory on chip is by establishing a good balance between redundancy and bit-cell sizes.

When the SRAM cells are examined at sub-threshold voltages, the write operation of the bit-cells is safe, but the read operation causes stability problems. In the read operation, the access speed causes the bit-cells to work slowly. The discharge current of the bit-cells can be fixed by using appropriate device sizing and using the right size of the bit-line capacitances. These bit-line capacitances have a direct influence on the slope of the charging and discharging of the bit-lines. For 65nm technology, 1pF bit-line capacitance is a value that pull-up and pull-down networks can handle. During the read operation of the bit-cells, both of the bit-lines are needed to be pre-charged to supply voltage. The size of the width of pre-charging PMOS transistor affects the charging up speed of the bit-lines. Operating at sub-threshold, 4µm pre-charge width enables the necessary speed to charge up the bit-lines.

2.1 6T SRAM

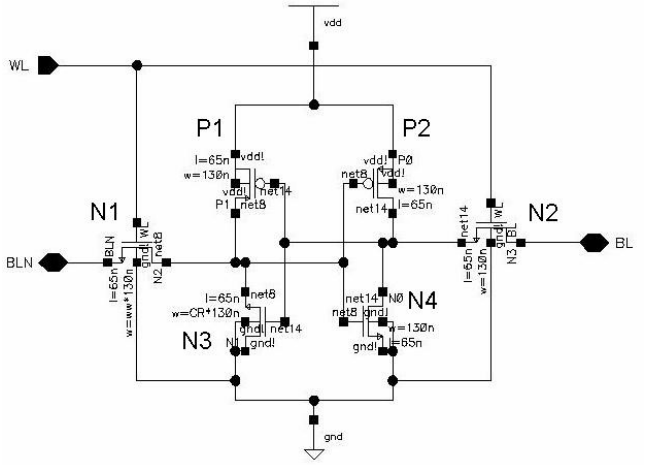

Figure 1. Schematic of 6T SRAM.

The six transistor SRAM cell is the most common design used in the industry. The transistors need to be sized in order to achieve the optimum results for both the read and the write operations. Sizing of the transistors affects the speed and the area directly. As width of the transistors increases, the access speed gets faster; the area increases and the noise margins become larger. When sizing in 6T SRAM, N3 would have to be wider than N1 in order to maintain read stability. As the N1 gets wider, the read static noise margin increases. Also the width of the P1 needs to be narrower than N1, so this derives the equation N3 > N1 > P1. For the write operation, the equation changes to N4 > N2 > P1 which determines the pull-up ratio of the bit-cell. The transient response shows the read operation of the bit-cell when the widths of N3 = 650nm, N1 260nm, and P1 = 130nm using 65nm technology (see

Figure 2). The slope of the bit-line is -825 kV per second which is approximately seven times faster than a non-scaled 6T SRAM's. The noise margins of the 6T SRAM can be calculated by method given in [4]. The difference of the two curves in Figure 3 would provide the diagonal dimensions of the largest square box. In this simulation, it would provide a SNM equaling 77.78 mV.

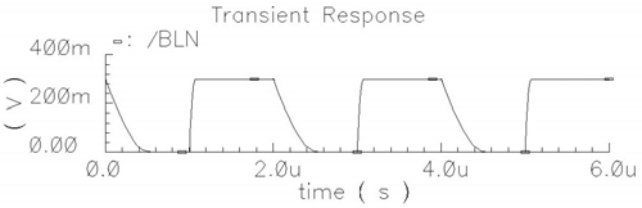

Figure 2. Transient response of 6T SRAM with sizing.

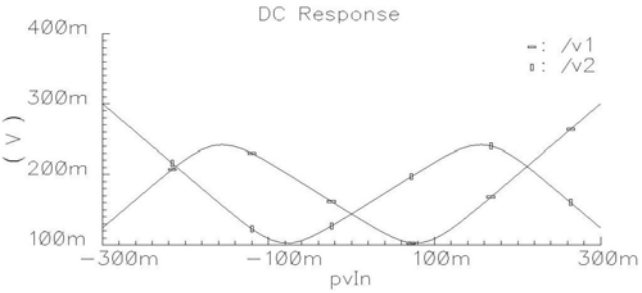

Figure 3. Static read noise margins of 6T SRAM in "u-v" domain.

2.2 8T SRAM

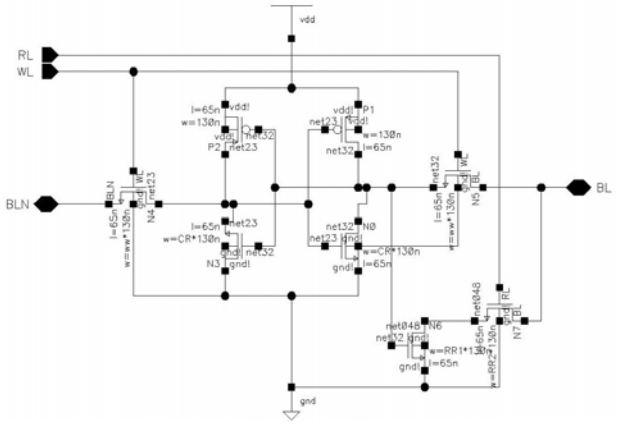

Figure 4. Schematic of 8T SRAM.

The eight transistors SRAM has two extra transistors when compared to the 6T SRAM [5]. These two transistors added to the pull-down network to help stabilize the read operation of the bit-cell. The sizing of these transistors affects the performance of the bit-cell directly. As the width increases, the speed increases, but the area gets bigger. When the widths of the transistors are 360 nm, the slope of the discharging bit-cell becomes -880 mV per second. The read SNM equals to 77.28 mV which is almost as same as the 6T SRAM's. The noise margins changes only when the two transistors are sized large. When compared to the 6T, the 8T is faster and has better stability, but larger in area.

2.3 4T SRAM

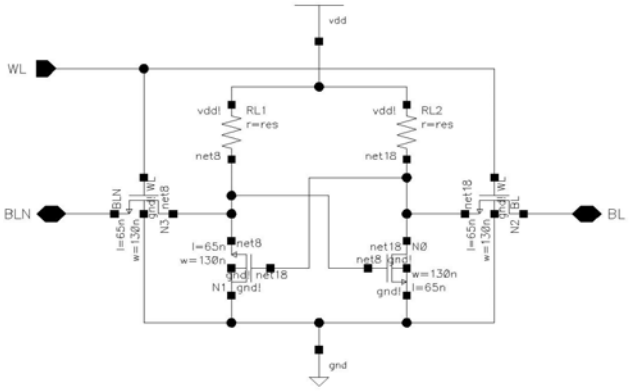

Figure 5. Schematic of 4T SRAM.

The four transistors SRAM, known as resistive-load SRAM, have a similar design to 6T SRAM. The only difference is the PMOS transistors are replaced by resistors. These resistors act as transistors in the bit-cells, but they help to reduce the cell size by approximately 1/3 compared to the 6T SRAM. The sizing of the resistors affects the performance of the SRAM. As the resistance increases, the cell works faster. When the resistor values are 600 kΩ, the discharging slope of the bit-cell equals to -710 mV per second. The biggest problem with the 4T SRAM is the production. The production becomes expensive, because of the extra steps needed in the manufacturing.

3. MODELING PROCESS VARIATION OF SUB-THRESHOLD SRAM

The preceding discussion is based on the assumption that all the parameters of a transistor are fixed and the functioning strictly follows this model. In actual fabrication process, however, the parameters of a transistor may vary from wafer to wafer (inter-die variations), or even between transistors on the same die (intra-die variations). Such variations will generate randomly distributed parameters, change the actual behavior of transistors and eventually introduce failures in memories.

Intra-die variations are caused by mismatches in parameters such as threshold voltage (V_T) and geometry (L, W). Threshold voltage fluctuation is especially critical to memory cells in sub-threshold operation because the current through a transistor observes an exponential relationship with the threshold voltage. V_T shift is considered to be a zero-mean Gaussian distribution, with standard deviation inversely proportional to square root of transistor area $(\sqrt{WL})$ [6] [7]. Larger transistors will exhibit less variation at the cost of extra die area.

In our following modeling of sub-threshold SRAM under process variation, we scale the sizes of transistors using a scaling factor. Monte Carlo simulation is performed on each size-set to compare the impact of sizes on its performance. Focus will be placed on two important metrics: Read-access Time and Static Noise Margin. They reflect the speed and reliability of a memory cell respectively.

Table 1. Transistor Sizes for running Monte Carlo Simulation

NMOS	PMOS	Pass NMOS	RR NMOS	Scaling Factor
650n*65n	130n*65n	260n*65n	390n*65n	1
1300n*65n	260n*65n	520n*65n	780n*65n	2
2600n*65n	520n*65n	1040n*65n	1560n*65n	4
6500n*65n	1300n*65n	2600n*65n	3900n*65n	10

3.1 Read-access Time

Read-access time is measured by the slope of the discharging curve (assume a 1 stored on the BL side of the back-to-back inverter, the BL will be pulled down to ground by the turned-on NMOS while reading).

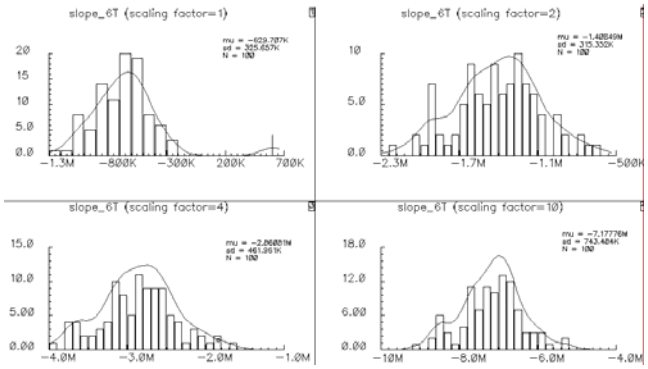

Figure 6. Read-access Speed (6T model)

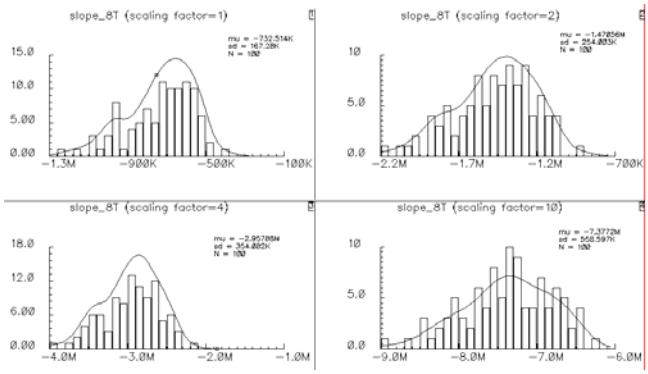

Figure 7. Read-access Speed (8T model)

As shown from the simulation results, larger sized transistors give larger mean value for discharging slope (thus, read-access speed), and the speedup is proportional to the scaling factor because the discharging current is proportional to W/L. On the other hand, the standard deviation of the slope also increases (it's not a linear function of V_T, although the standard deviation of V_T definitely decreases with larger transistor area). However, the deviation is does not proportionally increase as much as the mean value, which indicates the fact the percentage deviation $(\frac{\text{Standard Deviation}}{\text{Mean}})$ is decreasing.

The 8T models outperform the 6T ones in that they produce sleeper slope and exhibit less variation. Note in the result 6T model with scaling factor equal to 1, several data-points fall into the positive slope bin, which suggests a failure or malfunction. The occurrence of this situation is rarer in larger scaling factors.

3.2 Write-access Time

Write-access time is much smaller (faster than read-access by order of 2). This is because drivers that force BL/BLN to the desired values can be large during the read.

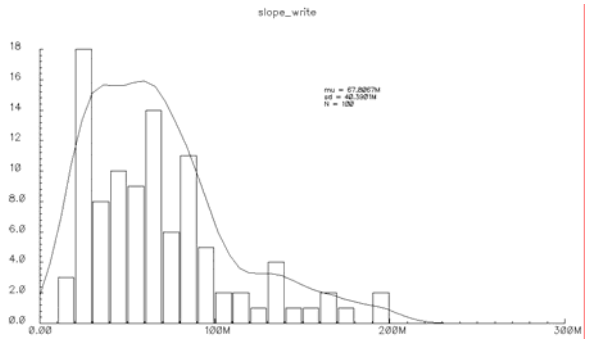

Figure 8. Write-access Speed

3.3 Static Noise Margin (SNM)

SNM is a measurement of the reliability of a memory cell. It is defined as the length of the side of the largest square that can be embedded inside the lobes of the butterfly curve [6]. We will focus out discussion on Read-SNM because the read operation charges up the internal node and severely degrades the SNM.

Process variation leads to serious problems in the SNM. Changes in V_T have a direct impact on the strengths of transistors. As a result, the VTC of a single transistor will shift left or right accordingly. This shift, in turn, changes the appearance of butterfly curve and will transform into a random fluctuation of SNM. Normal operation of a memory cells requires two stable operation points to maintain state. It is considered a failure if one of the two stable points is missing, which can happen due to these variations in V_T. The advantage of increasing transistor sizes is evident: larger Read-SNM and less standard deviation.

Data falling into negative SNM bins denote the cells that failed. For example: take the 6T model: when scaling factor equal to 1, the failure rate is 6% (6 fails out of 100), and 2% for scaling factor=2. Larger scaling factors further restrict failure rate to lower than 1%. An additional upside of larger transistors is a lowered failure rate (improved reliability). However, utilizing bigger transistors can create a larger than desired area overhead. Introducing redundant rows and columns may be a better solution to this issue.

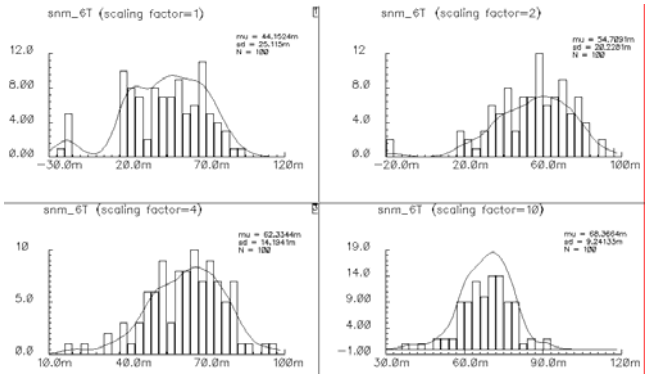

Figure 9. Read-access SNM (6T model)

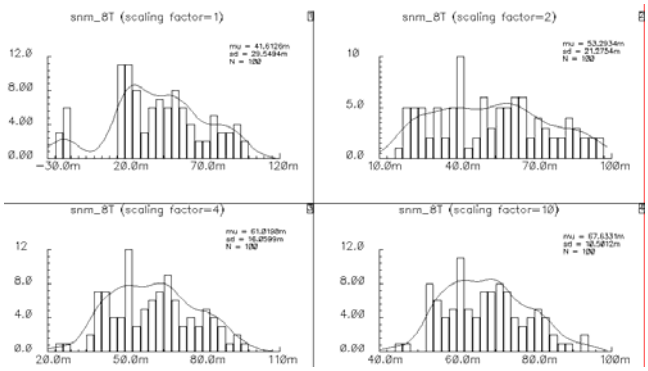

Figure 10. Read-access SNM (8T model)

3.4 Redundant Arrays

Redundant arrays are added to supplement the primary memory array. When any of the primary memory cells fail, the redundant ones can be used as substitute. In the preceding example, we can add 6 additional memory cells to recover from those failures. To be fail-safe, we still need one more (totaling 7) because these redundant cells are also susceptible to process variation. With just 7% area overhead, we can reduce the failure rate to an almost negligible degree while sizing the transistors to be twice larger (100% area overhead) still leaves us with 2% failure rate.

4. SUBTHRESHOLD SENSE AMPLIFERS

In addition to SRAM bit-cells, the operation of sense-amplifiers was also examined in the sub-threshold region. The sense amplifier is the device in an SRAM memory at the end of the bit-lines that detects a differential voltage between them before the low line has a chance to completely resolve. The advantage of the sense amplifier is in a great increase in access speed as well as a substantial power savings due to the fact that the low bit-line need not be pulled completely to GND.

Two classes of sense amplifiers were compared: the voltage-sense type and current-sense type. Each of the amplifiers was analyzed in for both speed and reliability, (quantified as resolution speed and input-referred offset voltage, respectively).

The voltage-sense amplifier is the most simple and most common sense amplifier found in SRAM memories in industry, and its schematic is shown in Figure 12 [8]. Simulations have shown that it does indeed work at sub-threshold voltages. It consists of nine transistors, forming a differential amplifier with cross-coupled inverter loads. The outermost PMOS transistors, P1 and P7, pre-charge the outputs to VDD when the enable is off. The bottom NMOS transistor, N8, forces the input transistors (N1 and N3) to amplify their differential voltage rather than their input voltage. Finally, the cross-coupled inverters force the amp to pull the differential outputs to the rails (i.e., VDD and GND) once a significant differential has been detected. The current-sense amplifier, on the other hand, was found to not work under sub-threshold conditions. A small signal analysis suggested that the current flowing through the transistors used to create a bias voltage at the input overpowered the differential currents created by the bit-cell.

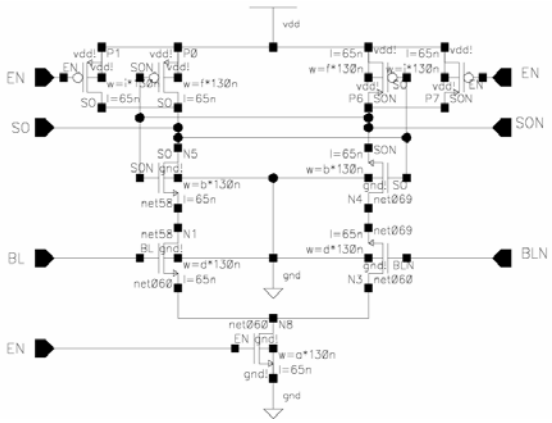

Figure 12. Schematic of a Voltage-Sense Amplifier

4.1 Analysis of Sense Amplifier Performance

Each amplifier was analyzed for both speed and reliability. These two abstract concepts were quantified by two metrics: resolution speed and input-referred offset voltage, respectively. Due to the effects of process variations, both of these metrics were measured using Monte Carlo simulations.

Resolution speed (RT) is a metric that measures the time it takes the amplifier to resolve once a significant differential voltage has occurred on the bit-lines. Physically, it is the maximum of the rise and the fall times of the outputs after the enable has been triggered. It will change based on the rate at which the voltage offset forms (and thus based on which bit-cell design from Section 2 was used); the sense amps in Section 4 were measured using the 6T bitcell. Figure 13 shows that the voltage-sense amplifier had a mean resolution time of 8.21ns with a standard deviation of 2.44ns.

Input-referred-offset-voltage (IROV) is defined to be the differential voltage at which the bit-lines must be at in order to ensure that the outputs resolve correctly. IROV is caused by process-mismatch variations in V_T of the input transistors in the sense amp. This mismatch acts like an error in the input differential voltage. Figure 14 shows a plot of IROV for the voltage-sense amplifier. As to be expected, it is approximately centered around zero and thus zero-mean. The plot also suggests that to be completely safe the amp must use the value at three standard deviations from the mean, approximately 25mV. Figure 15, a plot of the absolute IROV, better illustrates this point.

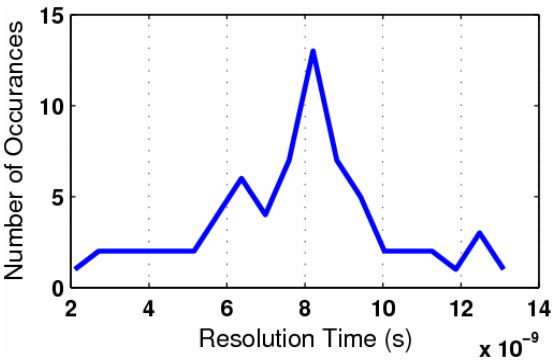

Figure 13. Resolution Time of a Voltage Sense Amplifier Under Influence of Process Variations

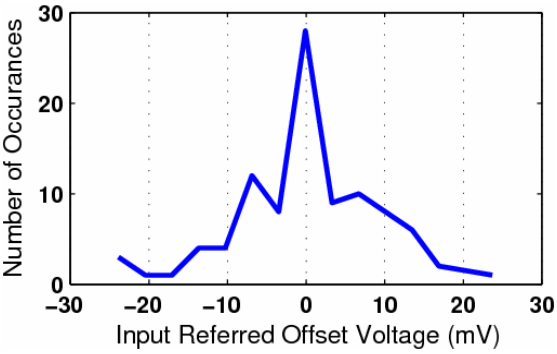

Figure 14. Plot of Input Referred Offset Voltage

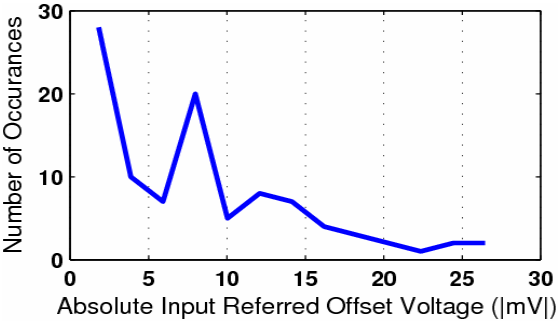

Figure15. Plot of the Absolute Value of IROV

4.2 Improving Sense Amplifier Performance

Analysis of the voltage-sense and strong-arm amps has identified two techniques to increase resolution speed: body-biasing PMOS transistors to increase output-current, and adding a regulated-cascode stage to increase output-impedance. The combined effect of these techniques has demonstrated an increase in resolution speed by as much as 4x (from about 8ns to 2ns).

Body-biasing the PMOS transistors of the sense amp means to connect their body to GND instead of VDD, lowering the transistors' threshold voltage due to the body-effect. This will increase the output current of the transistor and thus increase its intrinsic gain. [9] The effect of body-biasing on the voltage-sense amplifier was that it improved the resolution time by about 35% and yet almost doubled the IROV. The IROV was also nonzero-mean for the body-biased amp.

The regulated-cascode stage improves output-impedance by routing the amp's load (in this case, the cross-coupled inverters) through another differential amplifier (with gain A), thus multiplying the output-impedance of the sense amp by A. [10]. The regulated cascode stage doubled the resolution time of the body-biased amp, but also further increased the IROV by 20mV.

4.3 Sense Amplifier Conclusions

Although the body-bias and regulated-cascode techniques did improve the resolution time of the voltage-sense amplifier, they also negatively affect the input-referred-offset voltage. Therefore, in order to determine the optimal amp scheme one must look at the whether the bit-line takes longer than 3ns to drop 20mV (as is the difference between the different schemes).

Table 2. Table of Sense Amp Metrics, Approximate Values

Sense Amplifier	Mean RT	Mean IROV at 3σ
Voltage-Sense	8ns	25mV
Voltage-Sense w/ BB	5ns	45mV
Voltage-Sense w/ RC	2ns	65mV

5. REFERENCES

[1] V. Degalahal, N. Vijaykrishnan, and M. J. Irwin. *Analyzing Soft Errors in Leakage Optimized SRAM Design.* Proc. Of 16th International Conference on VLSI Design, pp. 227-233, Jan. 2003.

[2] Stanley E. Schuster. *Multiple Word/Bit Line Redundancy for Semiconductor Memories.* IEEE J. Solid-State Circuits, vol. SC-13, no. 5, pp. 698-703, Oct. 1978.

[3] E. Seevinck, F. List, and J. Lohstroh. *Static noise margin analysis of MOS SRAM cells.* IEEE J. Solid-State Circuits, vol. SC-22, no. 5, pp. 748-754, Oct. 1987.

[4] Raymond Heald and Ping Wang. *Variability in Sub-100nm SRAM Designs.* ICCAD, pp. 347-352, 2004.

[5] F. Lai, et al. *A New Design Methodology for Multiport SRAM Cell.* IEEE Transactions on Circuits and Systems, Part I: Fundamental Theory and Applications, vol. 41, no. 11, pp. 677-685, Nov. 1994.

[6] Calhoun, B. H. et al *Static Noise Margin Variation for Sub-threshold SRAM in 65-nm CMOS.* IEEE Journal of Solid-State Circuits, VOL. 41, NO. 7, July 2006

[7] Chen, Qikai et al *Modeling and Testing of SRAM for New Failure Mechanisms due to Process Variations in Nanoscale CMOS.* Proceedings of the 23rd IEEE VLSI Test Symposium (VTS'05)

[8] Wicht, B.; Nirschl, T.; et al.; *Yield and speed optimization of a latch-type voltage sense amplifier*; Solid-State Circuits, IEEE Journal of Volume 39, Issue 7, July 2004 Page(s):1148 - 1158

[9] Narendra, S, et al; *Ultra-low voltage circuits and processor in 180nm to 90nm technologies with a swapped-body biasing technique*; Solid-State Circuits Conference, 2004. Digest of Technical Papers. ISSCC. 2004 IEEE International 15-19 Feb. 2004 Page(s):156 - 518 Vol.1

[10] Nissinen, J.; Kostamovaara, J.; *Fully Differential, Regulated Cascode Amplifier*; Electrotechnical Conference, 2006. MELECON 2006. IEEE Mediterranean 16-19 May 2006 Page(s):51 - 54